Title
Adjacent and Non-Adjacent Word Contexts Both Predict Age of Acquisition of English Words: A Distributional Corpus Analysis of Child-Directed Speech

Journal
Cognitive Science, 44(11)

ISSN
0364-0213

Authors
Chang, Lucas M
Deák, Gedeon O

Publication Date
2020-11-01

DOI
10.1111/cogs.12899

Peer reviewed

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 44 (2020) e12899

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12899

Adjacent and Non-Adjacent Word Contexts Both Predict Age of Acquisition of English Words: A Distributional Corpus Analysis of Child-Directed Speech

Lucas M. Chang, Gedeon O. Deák

University of California, San Diego

Received 21 August 2018; received in revised form 27 July 2020; accepted 4 August 2020

Abstract

Children show a remarkable degree of consistency in learning some words earlier than others. What patterns of word usage predict variations among words in age of acquisition? We use distributional analysis of a naturalistic corpus of child-directed speech to create quantitative features representing natural variability in word contexts. We evaluate two sets of features: One set is generated from the distribution of words into frames defined by the two adjacent words. These features primarily encode syntactic aspects of word usage. The other set is generated from non-adjacent co-occurrences between words. These features encode complementary thematic aspects of word usage. Regression models using these distributional features to predict age of acquisition of 656 early-acquired English words indicate that both types of features improve predictions over simpler models based on frequency and appearance in salient or simple utterance contexts. Syntactic features were stronger predictors of children's production than comprehension, whereas thematic features were stronger predictors of comprehension. Overall, earlier acquisition was predicted by features representing frames that select for nouns and verbs, and by thematic content related to food and face-to-face play topics; later acquisition was predicted by features representing frames that select for pronouns and question words, and by content related to narratives and object play.

Keywords: Language acquisition; Language input; Vocabulary; Word learning; Syntax; Distributional semantics; Age of acquisition; Statistical learning

1. Introduction

Infants' linguistic, social, and physical environment provides a wide array of features that might support early word learning. How do the myriad cues and correlations in the

Correspondence should be sent to Lucas M. Chang, University of California, 9500 Gillman Drive, Mail Code 0515, San Diego, San Diego, CA 92093. E-mail: l8chang@ucsd.edu

world combine to determine which words are learned, and when? Laboratory studies have demonstrated, at various ages and situations, influences of factors including dominance of the word's referent in the child's visual field (Yu & Smith, 2012), speech–motion synchrony (Gogate & Bahrick, 1998), children's biases to attend preferentially to relevant cues such as object shape (Landau, Smith, & Jones, 1988), discourse context (Horowitz & Frank, 2015; Sullivan & Barner, 2016), and the distribution of exposures over time (Childers & Tomasello, 2002; Vlach & Johnson, 2013). However, laboratory studies typically involve mapping novel nouns to novel toys during a single experimental session, which is not representative of normal word acquisition. During typical language acquisition, early-learned words represent diverse syntactic and semantic types and are experienced repeatedly over weeks or months. To understand the role of these experiences in word learning, researchers have related the normative age of acquisition (AoA) of words to measures of their frequency and usage patterns in infant-directed speech, as reviewed below.

1.1. Predicting age of acquisition

Recently, large databases of infants' vocabularies collected using the MacArthur–Bates Communicative Development Inventory (CDI; Fenson, March man, Thal, Dale, & Reznick, 2007) have made it easier to study the normative AoA of individual words (Frank, Braginsky, Yurovsky, & Marchman, 2017). Infant-directed speech corpora such as those in CHILDES (MacWhinney, 2014) make it possible to derive predictors from real-world usage patterns. Using these data, Braginsky et al. found that normative ages of comprehension and production in seven languages are predicted by high word frequency, low mean length of utterance (MLU), frequent appearance in isolated or utterance-final positions, high concreteness ratings, and high "babiness" ratings by adults—that is, relevance of a word to babies (Braginsky, Yurovsky, Marchman, & Frank, 2016, 2019). Machine-extracted prosodic features also predict age of comprehension (Frermann & Frank, 2017). Swingley and Humphrey (2018) extended these analyses to predict individual differences. By pairing individual children's CDIs with samples of their parents' speech, they found that word frequency, frequency of occurrence in isolation, shorter utterance length, and longer spoken duration predicted both comprehension and production, and these predictors were stronger for matched mother–infant pairs than for randomized pairings of mothers and infants. Another study used dense data collected for a single child, finding that word production was predicted by frequency of a word in the input and by distinctiveness of a word's spatial, temporal, and topic distribution (Roy, Frank, DeCamp, Miller, & Roy, 2015).

The above factors illustrate that words can be easier or harder to learn in a few distinct ways: Some predictors relate to the sheer number of opportunities to learn the word, some relate to the conceptual accessibility of the word meanings for young children, and some relate to the position of words within a linguistic context. Within the last category, only a few simple predictors have been evaluated. These show that simplicity of the sentential context, distinctiveness of contexts, and placement at utterance boundaries, all positively predict word learning. However, data-driven distributional representations of word usage patterns have not been evaluated as predictors of AoA, even though distributional models have been proposed as mechanisms for infants to learn about word class (Clair, Monaghan, & Christiansen, 2010; Monaghan, Chater, & Christiansen, 2005) and verb semantics (Laakso & Smith, 2007), and can successfully represent semantic similarity and categorical structure when trained on infant-directed speech (Huebner & Willits, 2018).

The goal of the current study is to investigate whether, and how, distributional properties of words contribute to predicting AoA. Distributional features are derived from co-occurrences among words or word sequences, but might reflect, to varying degrees, different types of information including syntactic, thematic, and taxonomic relations among words (Huebner & Willits, 2018). Therefore, a second goal of the study is to design and extract a set of distributional features from real-world language use that segregates information types, so that the contributions of these types to lexical development can be evaluated separately. One simple way to segregate distributional features into two complementary streams is by tracking distributional statistics at different scales. That is, a word's distribution can be described both in terms of *adjacency relations* such as transition probabilities or simple constructions across successive words, and in terms of *non-adjacent co-occurrence* with other words in a wider scope. We expect this simple distinction to correlate with some higher level conceptual boundaries: Specifically, we expect that adjacency relations will primarily capture word class and syntactic information, whereas non-adjacent co-occurrence will primarily reflect thematic information. In the next section, we review the strengths and limitations of existing methods of generating distributional semantic representations of words.

1.2. Distributional word representations

General-purpose distributional lexical representations have a history of effective use in natural language processing, and therefore serve as a starting point for the design of developmentally relevant distributional representations. Three main approaches can be distinguished. One family of models represents words by their distribution across large-scale contextual units. These contexts can be defined by document boundaries in the training corpus, as in latent semantic analysis (Dumais, 2004), or they may represent topics inferred by a generative model, as in latent Dirichlet allocation (Blei, Ng, & Jordan, 2003).

A second group of models represents words by their co-occurrence rates with other words. This approach was introduced as the Hyperspace Analogue to Language (Lund & Burgess, 1996), and statistical refinements such as COALS (Rohde, Gonnerman, & Plaut, 2004). Recently, new levels of performance on large datasets have been achieved by recasting this approach as a neural network, as in Skip-gram (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013).

Finally, a third group of models learns to predict sequences of words, such that its internal states after training can be interpreted as word representations. This group includes recurrent neural networks (Elman, 1990) and their refinements such as LSTM (Hochreiter & Schmidhuber, 1997), Gated Recurrent Unit (Chung, Gulcehre, Cho, & Bengio, 2014), and Delta-RNN (Ororbia, Alexander, Mikolov, & Reitter, 2017).

Despite their successes, existing distributional models suffer from a few drawbacks. Models that represent words as points in a vector space have trouble accounting for the fact that pairwise word similarity is not a valid metric: Word A may be similar to word B, and word B may be similar to word C in an unrelated way, without implying that words A and C are at all similar (e.g., Barclay, Bransford, Franks, McCarreli, & Nitsch, 1974; Griffiths, Steyvers, & Tenenbaum, 2007; Medin & Shoben, 1988). Another potential problem is that computational models often achieve good performance by processing vast amounts of data at arbitrary timescales and memory loads, which are unrealistic models of human infants' learning, and might be counterproductive for certain kinds of learning (Elman, 1993; Newport, 1990; Phillips & Pearl, 2015).

Analyses of corpora of child-directed speech have demonstrated that children could potentially discover linguistic regularities by tracking the distributions of words. In particular, rich information about syntax is encoded in the distribution of words into contexts defined by the immediately adjacent words. For instance, many *frequent frames,* consisting of pairs of flanking words for a target word, are strongly selective for nouns and verbs, and have been proposed to form the basis for early syntactic learning (Mintz, 2003). Frames are an attractive measure for predicting AoA of specific words for several reasons. One is that they can readily be detected in corpora of child-directed speech. Another is that they encode information about word class and usage without requiring advanced syntactic knowledge. Another is that frames might support word learning in multiple ways. First, frequent frames likely facilitate word segmentation by demarcating the boundaries of the framed word, increasing the probability that it will be further processed. Second, frames help children infer word meaning by embedding words in familiar and meaningful constructions (see Gleitman, 1990; Goldberg, 2003). Third, frequent frames might facilitate production if infants acquire productive language in a construction-specific manner (Cameron-Faulkner, Lieven, & Tomasello, 2003; Tomasello, 2000). That is, if producing a word depends not only on inferring its meaning but also on having a construction in which it is known to occur, then the words that appear most often in the earliest acquired productive constructions are likely to be produced first. Finally, appearing in one or a few specific frames might be a proxy for a word's syntactic class.

In sum, features derived from word frame co-occurrence statistics merit closer study because frames are thought to be relevant for infant learning, they address open questions about interactions between syntactic and lexical learning, and they can be derived exclusively from the words immediately adjacent to each target word.

In contrast to adjacency relations such as frames, it has been noted that models based on non-adjacent word-co-occurrences, such as Skip-gram, tend to be biased toward thematic rather than syntactic or taxonomic information (Huebner & Willits, 2018). To further emphasize thematic information at the expense of syntax, a model based on word co-occurrence can be trained using a sliding window that counts co-occurrences between nearby words, but *not* adjacent ones. The representation modified in this way is a simple, associative model of coarse-grained distributional information whose input is complementary to that of frames.

Because co-occurrences are derived from bottom-up distributional statistics in infant-directed speech, they reflect the structure of naturalistic language environments more directly than theory-based measures of grammatical class and word meaning that have been used to predict acquisition (e.g., Braginsky, Yurovsky, Marchman, & Frank, 2016, 2019).

Thus, taken together, we expect frames and non-adjacent co-occurrences to reflect primarily syntactic and thematic aspects of word meaning, respectively, and to generate a rich set of quantitative features that jointly predict AoA better than either level of distributional information alone.

1.3. Current study

The current study addresses how distributional properties of words predict AoA of individual words. We derive quantitative features from a large corpus using two types of information: *Syntactic features* are derived from distribution of words in frames, and *thematic features* are derived from non-adjacent co-occurrences between words. We first demonstrate that these two types of features succeed in capturing different levels of structure in the lexicon, both qualitatively and by quantifying the extent to which they distinguish between syntactic and thematic word classes as defined by the groupings of words on the CDI infant vocabulary checklist. Then, we evaluate how well normative AoA of words is predicted by the two types of distributional features, using CDI production norms as the primary measure of AoA. We also used the features to predict CDI comprehension norms to increase generality and comparability with previous studies. We fit models predicting AoA using a set of simple word use metrics that have been previously described (i.e., frequency, MLU, final frequency, solo frequency; Braginsky, Yurovsky, Marchman, & Frank, 2016). We then investigate how these models are improved by including the syntactic and/or thematic distributional features.

2. Methods

2.1. Corpus

The corpus was constructed by downloading all American English transcripts from CHILDES (MacWhinney, 2014) and selecting those containing speech by primary caregivers directed to typically developing children aged 48 months or younger. Special codes indicating non-words were removed.[1] All punctuation was removed except for utterance boundary tokens. Dialectal, spelling, word segmentation, contractions, and transcription-style variants were standardized. Whenever possible, we attempted to standardize variants so that each CDI item corresponded to a single word type in the corpus. Finally, we removed common inflections and contractions (*-ing, -ed, -s, 'll, 've, 'm, 're, -n't*), leaving word lemmas. The resulting corpus contained 1,050,868 utterances and

4,533,900 tokens from 3,794 transcribed sessions (range: 2–8,648 tokens per recording session).

2.2. Age of production

CDI data were obtained from Wordbank (Frank et al., 2017). American English words and gestures (age 8–18) and words and sentences (age 16–30) were both used. For each age, the proportion of children producing the word was calculated. We then fit a logistic curve to the proportion, constraining the proportion to approach 1 (extrapolated). The age at which this curve crossed .5 was considered the age of production for the word. For model fitting, values were normalized to have zero mean and unit variance. Three items were dropped because they consisted of family-specific names; nine were dropped because they were phrases that could not be segmented as distinct units in the corpus; and two were dropped because they are sex-specific body parts and were very infrequent in the corpus. This left scores for 656 words.

2.3. Age of comprehension

Comprehension scores were calculated the same way as production scores, except that only the shorter words and gestures form includes a comprehension checklist (see Fenson et al., 2007). Thus, data were available for a smaller set of words and ages. In addition to the items that were dropped from the production dataset, logistic curves could not be fit to four words (brother, sister, mommy, daddy) because no increasing trend was present over the range 8–18 months. This resulted in scores for 383 words.

2.4. Syntactic features

Frequent frames were defined as pairs of words (including the utterance boundary marker, so that the framed word could be in utterance-initial or -final position) that co-occurred, separated by one word, at least 1,000 times in the corpus. This cutoff was chosen to eliminate most frames involving content words specific to a theme or activity, and those that occurred too infrequently to be familiar to most children. This yielded 439 frequent frames. All other frames were collapsed into a single "other" frame, representing rare or unknown frames. For each target word, we then constructed a vector of the number of times the word occurred in each frame. Thus, 656 words' occurrences in 439 frames were counted, yielding a 656×439 matrix. Values were first normalized so that each word's features (i.e., frame occurrences) summed to 1, and then, these features were scaled to have a mean of zero and variance of 1. Each row of the resulting matrix thus consisted of a 439-dimensional representation of a word, where each dimension represents its tendency to occur in a specific frame.

To reduce the dimensionality of this matrix into a smaller set of cohesive features, we applied principal components analysis (PCA). This operation produces a set of abstract features representing combinations of frames that best explain the variability in the data, with the constraint that they are uncorrelated with each other. For instance, all frames that allow mainly nouns are likely to contribute to a single principal component (PC) that represents, roughly, the affinity of a word for “noun” contexts. The resulting PCs are ordered by amount of variance explained, and by selecting the first few PCs, we obtain a low-dimensional set of syntactic features that capture the most divergent and systematic differences in the typical frame contexts of different words. Finally, to improve robustness of regressions, PCs were winsorized, clipping values more extreme than 10 times the mean absolute deviation from the median. We focus on the first 10 PCs for analysis. This process and the process described below are summarized schematically in Fig. 1.

2.5. *Thematic features*

Thematic (i.e., non-adjacent) features were generated using a version of the COALS model, based on correlations between words (Rohde et al., 2004). This model was chosen for its simplicity and lack of domain- or task-specific assumptions, and because it represents co-occurrences between words as statistical associations, in accordance with statistical learning theories of language acquisition (Erickson & Thiessen, 2015; Romberg & Saffran, 2010; Smith, Suanda, & Yu, 2014). Context words were defined as any word that occurred at least 1,000 times in the corpus (this cutoff eliminated rare context words that

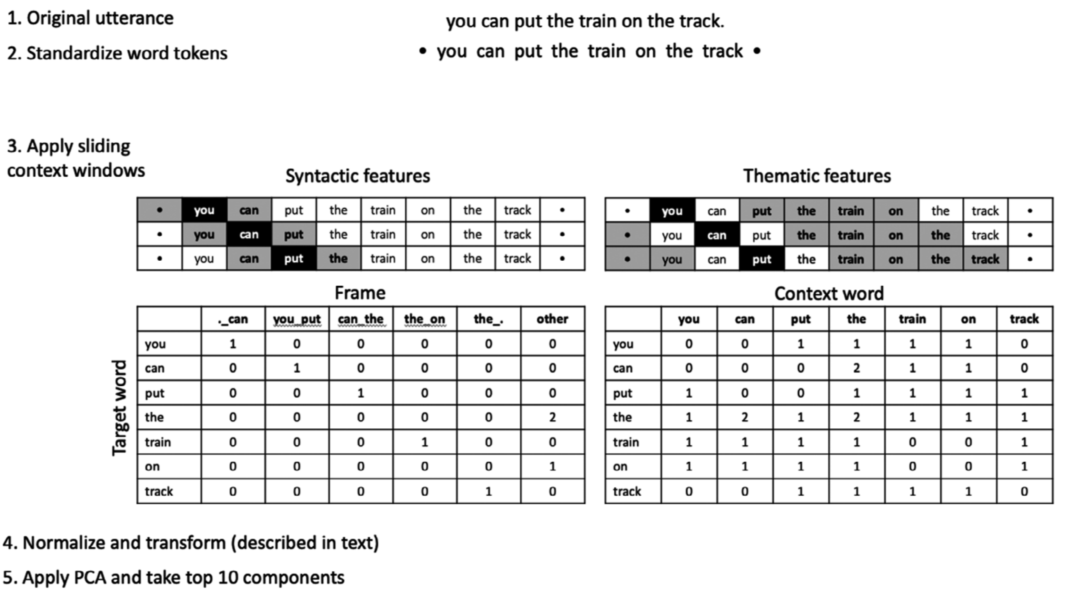

Fig. 1. Example computation of distributional features. Starting with an example utterance, we first standardize word variants, remove inflections, and insert utterance boundary tokens. We then apply two sliding windows to the input: for the syntactic features, we count the number of times each word (shaded in black) appears in the frame defined by the two surrounding tokens (gray). For the thematic features, we count the number of times each word (black) appears within five tokens of, but not adjacent to, each context word (gray). The resulting matrices are then transformed as described in the text, and the top 10 principal components of each matrix are the final features.

were unlikely to be familiar to children). We thus identified 1422 distinct context words. For each target word, we counted the number of times it co-occurred with each context word, using a sliding window to count co-occurrences that were within five flanking words but not adjacent. This resulted in a 656 $\times$ 1,422 matrix. Co-occurrence frequencies were then normalized to represent correlations, and negative values were set to zero (because the model is only interested in finding regular associations between words). Finally, values were square-root transformed. (As Rohde et al. [2004] note, this increases the relative weight of weak associations between words, thereby increasing sensitivity to patterns in a limited dataset, and it puts the values on a more interpretable scale.) Each row of the matrix thus consists of a 1,422-dimensional representation of a word, where each dimension represents its tendency to co-occur non-adjacently with a specific context word.

As above, we applied PCA to produce abstract features representing combinations of context words that explain the most variability among target words, and PCs were winsorized in the same way. This produced a set of 10 thematic features that are comparable in structure, but complementary in content, to the syntactic features.

2.6. Other metrics

Following Braginsky et al. (2016), we computed frequency, MLU, solo frequency, and final frequency. Frequency was calculated as the logarithm of the number of times each word appeared in the corpus. Stemming and standardization described above ensured that this frequency represents that of the word overall and not just the specific form listed in the CDI. MLU was calculated as the mean number of words in utterances containing each word, where utterance boundaries were defined by the original transcription, and did not always correspond to grammatical sentences. Solo frequency and final frequency were calculated by taking the logarithm of the number of times each word appeared in a one-word utterance or as the last word of an utterance, respectively, and finally taking the residual with respect to log frequency.

2.7. Evaluation

We first evaluate the extent to which syntactic and thematic features encode distinct information. We investigate this qualitatively by reporting the frames/context words that contribute most to the top PCs, and the words with the highest and lowest values on each PC. We then quantitatively evaluate the extent to which each PC distinguishes among either syntactic or thematic word categories, as defined by the groupings on the CDI, and confirm that the syntactic and thematic distributional features differentiate the syntactic and thematic word groups, as hypothesized.

We then fit multiple linear regression models predicting both age of production and age of comprehension for the target words. First, a baseline model included frequency, MLU, solo frequency, and final frequency. Next, a full model included all features as predictors. Finally, we fit three models with, respectively, the baseline and syntactic features, the baseline and thematic features, and the syntactic and thematic features (i.e., leaving out one feature set at a time). Models were evaluated in terms of R^2 and adjusted R^2, and significance for each predictor set was evaluated using likelihood ratio tests between the full model and the model without the predictors. Finally, predictive robustness of the full model was evaluated using 10-fold cross-validation to estimate the root mean squared error (RMSE) for predicted AoA of words not used to fit the model.

We additionally repeated the analysis with different values (increasing or decreasing by a factor of 2) for minimum feature frequency, number of PCs, and winsorization levels to ensure that the model results are not strongly dependent on the specific values chosen.

3. Results

We examined the top PCs of the syntactic (adjacent) and thematic (non-adjacent) distributional features. First, we examined scree plots to validate the number of PCs selected for each feature type (Fig. 2). These plots show the proportion of the total variance in the raw features that is explained by each additional syntactic and thematic PC. Inspection of the plots shows that the proportion of explained variance levels off around 5–10 PCs. Therefore, we conduct the main analyses using the first 10 PCs (models with 5 and 20 PCs were also evaluated as part of the robustness checks).

Next, we inspected the top PCs to determine whether they qualitatively capture mainly syntactic and thematic distinctions, respectively. Table 1 shows the words with lowest and highest values, and the frames with lowest and highest weights, for the first 10 syntactic PCs.

Examining the sets of syntactic word frames qualitatively, PC1 separates pronoun frames from verb frames, and PC2 separates nouns from verbs. Thus, the first two PCs encode distinctions among three major word classes. PC3 selects for modal verbs that appear between a pronoun and a verb. PC4 distinguishes between modal verbs and

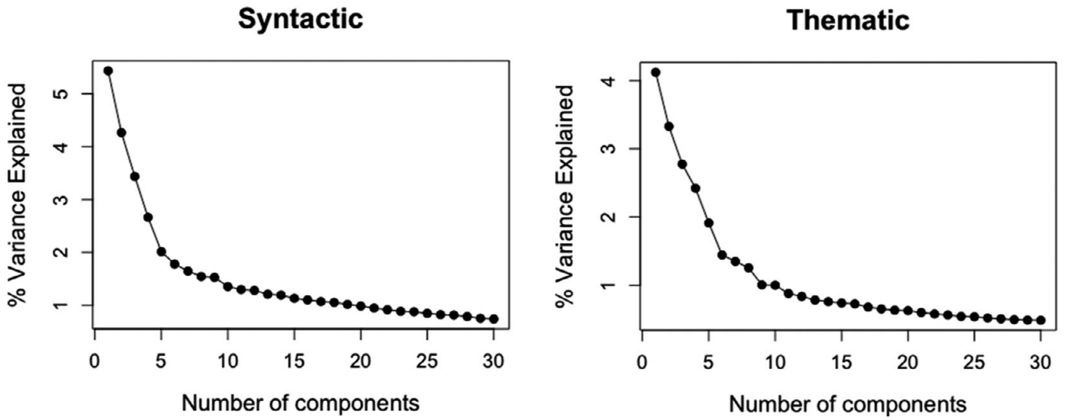

Fig. 2. Scree plots for PCA components.

Table 1
First 10 syntactic PCs

PC	Highest Words[a]	Lowest Words	Highest Weight Frames	Lowest Weight Frames	Production Coefficient[b]	Comprehension Coefficient
1	*read*	*you*	*you_the*	*did_do*	−0.58***	−0.45
	hear	*we*	*can_it*	*so_can*		
	fix	*they*	*to_it*	*._gotta*		
	put	*i*	*gonna_it*	*._could*		
2	*tights*	*hear*	*the_.*	*you_the*	−0.32*	0.30
	bead	*put*	*a_.*	*can_it*		
	block	*read*	*the_in*	*wanna_the*		
	cheerio	*fix*	*your_.*	*to_that*		
3	*you*	*wanna*	*can_put*	*you_put*	−0.01	0.07
	we	*gotta*	*do_see*	*you_get*		
	they	*gonna*	*are_gonna*	*you_take*		
	touch	*could*	*do_wanna*	*you_do*		
4	*you*	*what*	*do_see*	*._he*	−0.86***	−0.38
	wanna	*where*	*do_wanna*	*._she*		
	gotta	*why*	*do_think*	*._did*		
	have_to	*who*	*do_want*	*._do*		
5	*she*	*you*	*._was*	*can_say*	−0.34*	−0.24
	he	*are*	*._has*	*do_remember*		
	what	*did*	*._does*	*do_think*		
	who	*is*	*there_is*	*do_want*		
6	*she*	*you*	*is_.*	*the_.*	−0.04	−0.23
	he	*what*	*was_.*	*._do*		
	sick	*where*	*he_.*	*a_.*		
	stuck	*why*	*they_.*	*the_in*		
7	*we*	*what*	*oh_you*	*._do*	−0.40***	−0.78***
	no	*you*	*._they*	*know_that*		
	now	*how*	*._i*	*._is*		
	are	*why*	*._we*	*._did*		
8	*under*	*don't*	*it_the*	*you_.*	−0.29**	−0.26
	is	*what*	*go_the*	*._do*		
	in	*why*	*go_there*	*you_like*		
	over	*draw*	*it_there*	*wanna_.*		
9	*wanna*	*don't*	*._what*	*what_you*	−0.24*	−0.76**
	now	*are*	*._let*	*where_you*		
	you	*did*	*._how*	*what_we*		
	here	*we*	*._where*	*where_the*		
10	*show*	*build*	*._me*	*we_a*	−0.11	0.01
	give	*draw*	*me_you*	*gonna_a*		
	help	*is*	*._mommy*	*you_a*		
	scare	*need*	*you_me*	*to_a*		

[a]Signs are arbitrary, so "highest" is chosen for the extreme that predicts lower age of production; [b]Coefficients are reported for the full model including baseline predictors and both sets of PCs. Significance is computed with likelihood ratio tests; *$p < .05$; **$p < .01$; ***$p < .001$.

question words. PC5 seems to encode the difference between *you* and other pronouns. PC6 selects for adjectives, and PC7 seems to encode both question words and transition frames where a word precedes a pronoun at the start of an utterance. PC8 selects for location words, PC9 selects for helping verbs, and PC10 distinguishes between verbs that take human or inanimate objects.

Similarly, Table 2 shows for the first 10 thematic PCs, the words with lowest and highest values, and the context words with lowest and highest weights. PC1 separates words that appear in food contexts from words that appear in narrative contexts (some of the extreme values occur for low-frequency object words that are not semantically related, but nonetheless associate with food-related context words). PC2 selects for the topic of animals; PC3 and PC4 collectively distinguish between object play contexts and talk about the past. PC5 separates object play from interpersonal play contexts. PC6 separates greetings and communication-related contexts from body-related contexts; PC7 separates object-describing contexts from a mix of food-related and movement-related context words; PC8 separates dressing and bathing from vehicle/toy play contexts. PC9 selects for colors, and PC10 seems to encode additional distinctions between animal-related and narrative contexts. Taken together, the syntactic features reflect salient aspects of word class, whereas the thematic features represent different topics and activities.

We next verified that the syntactic and thematic PCs quantitatively discriminated between different syntactic and thematic groups of words, respectively. For this purpose, we used the word categories on the CDI. We collapsed the categories into a subset that primarily reflects syntactic distinctions, and a subset that primarily reflects thematic distinctions. The "syntactic" categories were the following categories from the CDI: quantifiers, locations, helping verbs, connecting words, descriptive words, action words, pronouns, question words, and the 11 noun categories (combined into a single set). Time words, sounds, and games/routines were dropped because they were not syntactically homogeneous. The "thematic" categories were the 11 noun categories taken separately: vehicles, animals, body parts, food/drink, people, outside things, toys, furniture/rooms, household objects, places, and clothing. For each PC, we calculated the F-statistic reflecting the ratio of between-category and within-category variance for each category set. Fig. 3 shows the F-statistics for the first 10 PCs in each feature type for both category sets. All 10 frame features reflected syntactic categories significantly better than thematic categories (median = 55.3 for syntactic categories, median = 8.5 for thematic categories; Wilcoxon test, $p < .001$; rank-biserial correlation = 1), and 9 of 10 co-occurrence features reflected thematic categories significantly better than syntactic categories (median = 7.8 for syntactic categories, median = 27.9 for thematic categories; Wilcoxon test, $p < .05$, rank-biserial correlation = .75). This confirms that adjacent context better predicts syntactic class, whereas non-adjacent context better predicts thematic relations within a syntactic class (nouns).

Having established that the two feature sets differentiate distinct aspects of word semantics, we then used the distributional features to predict age of production and comprehension of words using linear regression models. For both production and comprehension, we evaluated a baseline model using frequency, MLU, solo frequency, and final

Table 2
First 10 thematic PCs

PC	Highest Words[a]	Lowest Words	Highest Weight Context Words	Lowest Weight Context Words	Production Coefficient[b]	Comprehension Coefficient
1	*ankle*	*the*	*want*	*and*	−0.51*	−.40
	melon	*he*	*yummy*	*day*		
	snowsuit	*and*	*spoon*	*vacation*		
	coke	*was*	*wipe*	*plan*		
2	*the*	*you*	*is*	*some*	−0.12	0.24
	who	*i*	*look*	*more*		
	quack	*eat*	*rabbit*	*i*		
	cat	*have*	*cat*	*alright*		
3	*was*	*put*	*had*	*oops*	−0.06	0.64**
	eat	*it*	*um*	*whoa*		
	day	*on*	*eat*	*whoops*		
	were	*up*	*was*	*pull*		
4	*to*	*that*	*went*	*mm*	−0.01	0.33
	when	*a*	*when*	*that*		
	day	*you*	*were*	*mhm*		
	time	*is*	*night*	*good*		
5	*hi*	*in*	*infant's name*	*of*	−0.48***	−0.36*
	say	*the*	*hi*	*we*		
	you	*put*	*aw*	*car*		
	kiss	*of*	*kiss*	*truck*		
6	*he*	*you*	*has*	*infant's name*	−0.005	−00.48**
	his	*bye*	*his*	*later*		
	and	*we*	*hurt*	*bye*		
	but	*wanna*	*still*	*hi*		
7	*the*	*that*	*jump*	*color*	−0.65***	−0.85***
	eat	*i*	*eat*	*pretty*		
	down	*these*	*milk*	*wear*		
	jump	*red*	*down*	*know*		
8	*put*	*it*	*pajama*	*it*	−0.33**	−0.65***
	and	*that*	*clothes*	*noise*		
	her	*car*	*dress*	*crash*		
	wash	*push*	*bath*	*break*		
9	*does*	*green*	*does*	*yellow*	−0.03	0.23
	he	*blue*	*pig*	*green*		
	think	*and*	*farmer*	*blue*		
	in	*red*	*think*	*red*		
10	*quack*	*did*	*rooster*	*drive*	−0.17	−0.19
	i	*you*	*cow*	*yesterday*		
	it	*he*	*moo*	*got*		
	moo	*at*	*sheep*	*head*		

[a]Signs are arbitrary, so "highest" is chosen for the extreme that predicts lower age of production; [b]Coefficients are reported for the full model including baseline predictors and both sets of PCs. Significance is computed with likelihood ratio tests; $*p < .05$; $**p < .01$; $***p < .001$.

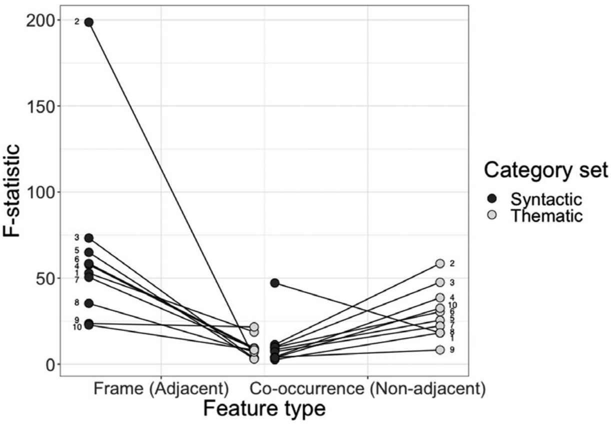

Fig. 3. Frame versus co-occurrence PCs segregate syntactic and thematic information. Each feature is represented as a line connecting its F-statistic with respect to the syntactic and thematic categories on the CDI. Downward sloping lines correspond to features that are more diagnostic of syntactic category, and upward sloping lines correspond to features that are more diagnostic of thematic category. Labels along sides indicate numbered PCs.

frequency; a full model including all features; and models with only the baseline and syntactic features; the baseline and thematic features; and the syntactic and thematic features, leaving out one set of features at a time. Models were evaluated using R^2, adjusted R^2, and root mean squared error (RMSE) under 10-fold cross-validation. Results are shown in Table 3.

For both comprehension and production, the full model performed substantially better than the baseline model (adjusted R^2 = .634 vs. .493 for production, adjusted R^2 = .460 vs. .242 for comprehension). Moreover, for both comprehension and production, the

Table 3
Model evaluation for different subsets of features

Model	R^2	Adjusted R^2	Cross-Validation RMSE
Production			
Full	.647	.634	2.49
Baseline	.496	.493	2.85
Syntactic + Thematic	.402	.383	3.35
Baseline + Thematic	.570	.560	2.69
Baseline + Syntactic	.603	.594	2.58
Comprehension			
Full	.494	.460	2.88
Baseline	.242	.234	3.31
Syntactic + Thematic	.366	.330	3.26
Baseline + Thematic	.417	.395	3.00
Baseline + Syntactic	.383	.360	3.03

model fit was significantly degraded when any of the three features sets was removed (likelihood ratio tests, ps < .001). For both comprehension and production, removing the baseline features caused the biggest increase in prediction error relative to the full model. However, the relative importance of syntactic and thematic features depended on the AoA measure: For comprehension, prediction error was higher for the model without thematic features (RMSE = 3.03 months) than for the model without syntactic features (RMSE = 3.00 months), whereas for production, prediction error was higher for the model without *syntactic* features (RMSE = 2.69 months overall, 2.71 for the subset of words that are on the comprehension form) than for the model without thematic features (RMSE = 2.58 months overall, 2.53 months for the comprehension form subset). This pattern suggests that syntactic features are relatively more important for production, whereas thematic features are more important for comprehension. Finally, error was higher for comprehension (RMSE = 2.88 months) than production (RMSE = 2.49 months overall, 2.50 on the comprehension form subset), though it is unclear whether this is due to weaker predictive relations or to noisier data due to the difficulty of assessing comprehension in infants.

Predictions of the full model of age of production and comprehension of individual words, plotted against actual AoA values for the production corpus and the comprehension corpus, are shown in Fig. 4.

To further characterize the contributions of individual syntactic and thematic features to the AoA values, we also report the estimated weights for each feature in the full models in Tables 1 and 2. These are visualized in Fig. 5 for production (top) and comprehension (bottom) data. Accuracy of estimation of the contribution of each individual feature is decreased when predictors are correlated; therefore, we calculated the variance inflation

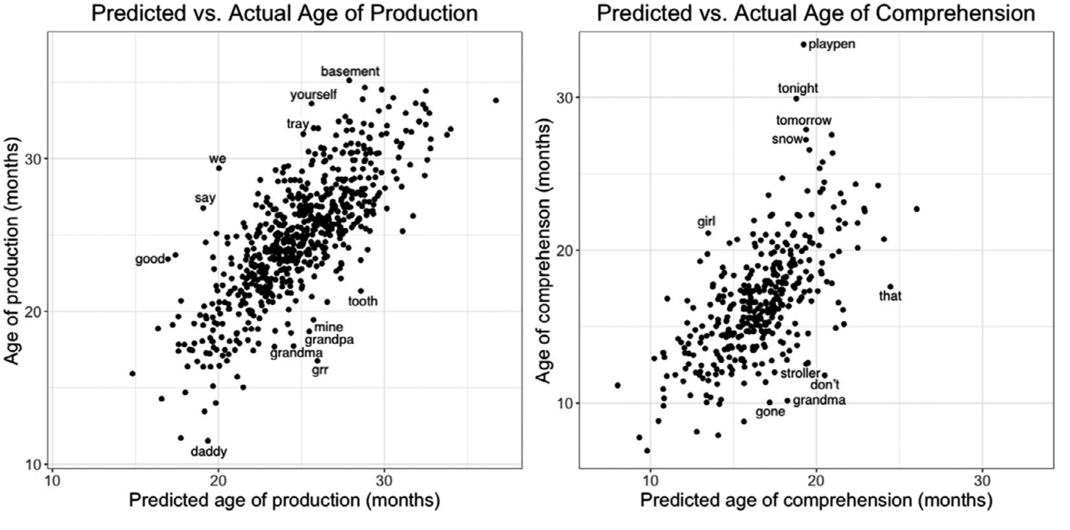

Fig. 4. Ten-fold cross-validation predicted versus actual AoA (left: production; right: comprehension) for all words.

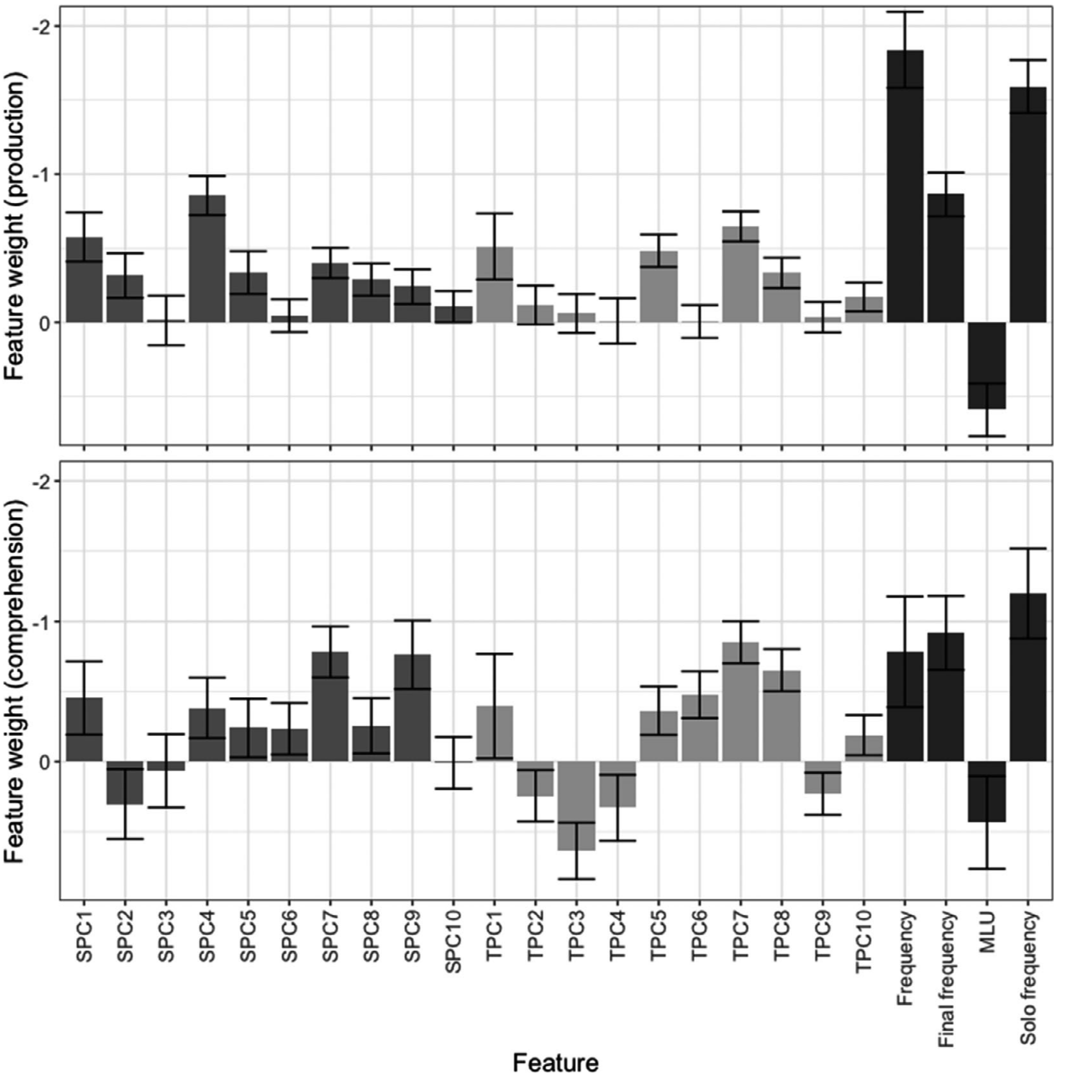

Fig. 5. Feature weights (top: production; bottom: comprehension). Distributional feature signs are chosen so that higher feature values predict earlier production; that is, negative weights (plotted up). Therefore, for comprehension weights, bars pointing up represent coefficients with the same sign as those for production, whereas bars pointing down represent coefficients with opposite sign. Error bars indicate standard errors of the coefficient estimates. SPC: syntactic principal component; TPC: thematic principal component.

factor for each feature in the full production and comprehension models (Tables S4 and S5). VIFs ranged from 1.05 to 7.41 in the production model, and from 1.22 to 6.06 in the comprehension model, indicating a moderate amount of multicollinearity.

Finally, to test whether our results were dependent on the specific choices for numerical parameters, we repeated all analyses with different values (increasing or decreasing by a factor of 2) for minimum feature frequency, number of PCs, and winsorization levels. The overall pattern of model fits was similar in all cases (Tables S1–S3). Performance was slightly degraded by increasing the minimum feature frequency, suggesting that less common items add noise to the PCs. Performance was also slightly degraded by increasing the winsorization threshold, suggesting that a few extreme PC values can obscure the information encoded in smaller differences among other words. Changing the number of PCs had different effects depending on the model: Decreasing the number of PCs from 10 to 5 degraded performance of all models, whereas increasing the number of PCS from 10 to 20 improved predictions slightly for production but degraded predictions for comprehension. However, increasing the number of PCs did not affect the pattern of relative performance among models. Thus, models using 10 PCs avoid overfitting while encoding most of the relevant information in the word usage distributions.

4. Discussion

In this paper, we introduce a novel distributional representation of word usage in child-directed speech. By deriving separate representations of words' distribution over frames and (non-adjacent) co-occurrences between words, we produce two sets of features that capture word usage patterns at different timescales. These features primarily encode syntactic and thematic information, respectively, which we confirm both by qualitative inspection of the feature content and by comparing feature values across syntactic and thematic word groups. Furthermore, the features are consistent with a simple model in which learners track the co-occurrences of frequent items in naturalistic input.

We then examine the degree to which each feature type predicted children's normative AoA of English words, compared to a baseline model that includes previously identified word usage features including MLU and three frequency-based measures (Braginsky et al., 2016). The baseline features, especially word frequency and solo frequency, were consistently among the strongest predictors in all models, suggesting that a word's distribution over usage contexts complements but does not supersede overall frequency and frequency in salient contexts. Nonetheless, for both production and comprehension, both distributional feature types predicted AoA over and above the baseline model. The syntactic and thematic feature types were complementary, in that a model that includes both feature types predicted AoA over and above either feature type alone. Together, syntactic and thematic distributional features explained 15% more variance in production AoA than the baseline model, and they explained 25% more variance in comprehension. These improvements were robust both to corrections for the number of parameters and to cross-validation.

Although comprehension and production showed similar patterns of incremental predictive utility for distributional features over and above frequency-based (baseline) features, age of comprehension was predicted less accurately overall than age of production. It is likely that comprehension data are inherently noisier due to limitations in parent report accuracy (Eriksson, Westerlund, & Berglund, 2002; Feldman et al., 2000; Oliver et al., 2003; Tomasello & Mervis, 1994). Syntactic features were a stronger predictor for production, whereas thematic features were a stronger predictor for comprehension. This suggests that availability of syntactic constructions (and related cognitive resources such as verbal working memory) might be a limiting factor in children's acquisition of words for production (Arnon & Clark, 2011). The relative influence of individual features was nevertheless mostly consistent between comprehension and production (see Fig. 5).

Different types of distributional features contributed to AoA in different ways. Frequency, solo frequency, and final frequency were consistently strong predictors, which suggest that a word's inherent salience in the input may contribute to learning rate, possibly in several ways—for example, solo frequency eliminates the problem of segmentation and suggests that the word can be meaningful without requiring the user to represent complex relations among multiple entities (Brent & Siskind, 2001). Consistently with this, concrete nouns and interactive social words (e.g., "hi") are prevalent among children's first words, as are frozen multi-word phrases that express a discrete meaning and segment as a unit (Bates et al., 1994; Lieven, Pine, & Barnes, 1992). Final frequency, in contrast, suggests a role of perceptual salience and/or working memory limitations in learning (Fernald & Mazzie, 1991).

Among the syntactic features, the strongest effects indicated that content words (nouns and verbs) were learned earlier than pronouns, and that question words were produced late, consistent with previous studies of vocabulary composition by word class (e.g., Bates et al., 1994; Caselli et al., 1995).

Among the thematic features, the strongest effects indicated that words used in food contexts were learned earlier than words used in narrative or descriptive contexts, and that words used in face-to-face interpersonal play were learned earlier than words used in object play. These patterns are consistent with the limited sentence complexity of early production, as well as the well-established developmental progression from dyadic to triadic and finally displaced reference (Adamson & Bakeman, 1991; de Barbaro, Johnson, Forster, & Deák, 2016; Morford & Goldin-Meadow, 1997; Sachs, 1983).

The predictive utility of these different types of distributional features derived from naturalistic caregiver speech, in addition to frequency-based features, suggests that distributional contextual information might play a role in infants' word learning. This identifies a gap in much of the research on word learning, which has focused more prominently on the referential transparency of object-naming events (McGillon et al., 2013; Medina, Snedeker, Trueswell, & Gleitman, 2011; Trueswell et al., 2016; Yu & Smith, 2012), social cues in naming (Frank, Tenenbaum, & Fernald, 2013), and children's biases about word (mostly noun) meanings (Deák, 2000; Markman, 1990). Thus far, these explanations of word learning have been complementary, in that the relevant factors are typically manipulated experimentally but are not assessed or estimated from naturalistic data. Conversely, frequency and distributional data can be computed from corpus data but not assessed experimentally in real-time learning. Furthermore, because experimental and distributional factors have typically been evaluated independently, it has been difficult to meaningfully compare the relative importance of each type of factor, alone or jointly. This methodological divide has obscured interactions among factors in word learning. However, new experimental tests of distributional factors identified in the current study and related work, in conjunction with dense recording of social cues and visual-speech co-occurrences enabled by new technology (e.g., Smith, Yu, Yoshida, & Fausey, 2015; Yurovsky, Smith, & Yu, 2013) should make it possible to estimate word learning as a function of multiple relevant factors in the local embodied social environment, as well as the more protracted accumulation of distributed linguistic patterns.

Our bottom-up approach to deriving word features aims to be less theory-laden and assumption-dependent than methods that attempt to characterize word syntax or meaning directly. Moreover, by representing words based on their usage in actual child-directed speech, we can be more confident that the features reflect distributional patterns that are observable by infants and children. Of course, many other types of word features vary with syntactic class—for example, nouns tend to be more concrete than other words (Brysbaert, Warriner, & Kuperman, 2014). Words also vary in their degree of perceptual grounding. For example, early-acquired nouns tend to refer to objects that occur frequently in infants' visual environments (Clerkin, Hart, Rehg, Yu, & Smith, 2017), and object-naming events vary in their referential clarity (Cartmill et al., 2013). Concrete nouns tend to be used more often in the presence of their referent compared to similarly frequent abstract words (Bergelson & Swingley, 2013). Social cues such as caregivers' gaze and pointing also vary in frequency across different words and constructions (Chang & Deák, 2019; Mason, Kirkpatrick, Schwade, & Goldstein, 2018; Murphy, 1978), and words vary in their spatial and temporal distinctiveness (Roy et al., 2015).

In our view, it is unlikely that distributional features exert their influence directly and independently of other cues. Instead, they provide us with a way to identify those contexts and constructions that most reliably afford word learning. Future research might then determine how those events support learning. It is likely that no single information source dominates; rather, the current results are consistent with the possibility that multiple correlated features in different modalities and in the language input conspire to distinguish groups of words (Bhatt, Wilk, Hill, & Rovee-Collier, 2004; Sahni, Seidenberg, & Saffran, 2010; Yu & Ballard, 2007). Indeed, it might be the quantity of correlated features, rather than their specific type, that is most important. Thus, it might be possible to achieve comparable performance in predicting AoA using independent feature sets such as distributional, acoustic, visual, and other features. As more modalities are added, redundant predictors might converge to the same set of words that occur frequently in distinctive, salient, and referentially transparent contexts.

It is important to note that distributional features may be proxies for other correlated cues, so correlations between distributional features and AoA do not necessarily imply that children learn by tracking word context distributions. We designed our features to differentiate words by syntactic class and by thematic/activity contexts. Therefore, we cannot determine how much our effects depend on the distributional features we measured, as opposed to the associated syntactic and semantic features, which children might detect in other ways. However, distributional features might help to identify which syntactic and semantic distinctions matter for word learning, thereby contributing to, rather than detracting from, theories of word learning based on these types of abstract features. Future research might further distinguish the contributions of word distributions by computing distributional features for each child based on their individual language input, rather than combining data from many children. By focusing on the *differences* in distributional features between children, it would be possible to eliminate any confounding effect of word features, such as grammatical class or concreteness, that are common to all individuals in a language community.

By measuring individual differences in exposure to words in syntactic and thematic contexts, future work might also identify sources of individual differences in vocabulary size and composition. Whereas individual differences in exposure to words within referentially transparent and/or developmentally appropriate contexts predict variability in children's subsequent receptive and productive vocabulary (Bergelson & Aslin, 2017; McGillion et al., 2013; Rowe, 2013), individual differences in word distributions might be used to predict variations in the order of acquisition of individual words. It might also be possible to measure the degree to which a caregiver's speech style supports word learning by characterizing how often they use words in contexts that are most strongly associated with word learning. Future research could investigate whether such data-driven measures of speech "quality" can supplement or even outperform existing summary measures of caregivers' speech quantity and style in predicting children's vocabulary outcomes (Tamis-LeMonda, Kuchirko, & Song, 2014; Weisleder & Fernald, 2013).

The current study has several limitations. Our measures of caregiver word use and AoA data were taken from different samples of children, and it is not possible to fully characterize the differences between these samples. In addition, the caregiver corpora come from a biased sample of situations recorded by researchers, so it is not clear how faithfully the corpora represent natural input to English-learning infants. Conversely, the AoA estimates are flawed by virtue of their sole basis in parental reports, which have documented limitations as noted above. Another limitation is that the current study treats AoA only as a population average. It is not known whether different features increase the probability of acquisition for all words across all children equally, or if there are individual differences in style or trajectory of acquisition. Yet, it is suggestive that interindividual variability exists in the distribution of syntactic types in children's vocabulary (Bates et al., 1994; Kauschke & Hofmeister, 2002). Moreover, children's existing vocabulary predicts which new words will be acquired subsequently (Beckage & Colunga, 2013). Therefore, understanding AoA as a non-stationary function of different types of predictors might enable better description and prediction of this variability.

Another limitation related to the corpora is that we did not model differences in parental word usage as a function of children's age. Developmental changes in language input might induce associations between AoA and specific constructions or contexts. In addition, the types of contexts that predict learning might change with age as children become increasingly able to process complex and/or decontextualized language (Rowe, 2012, 2013). Finally, a further limitation of the generalizability of the results is that languages other than English might not have the same degree of correspondence between syntactic constructions and frames defined by adjacent words. Frames might be expected to be less informative in languages that make greater use of inflectional morphemes and non-adjacent dependencies such as agreement, and less use of word order. Thus, although semantic and frequency predictors of AoA have been found to be largely consistent in a variety of languages (Braginsky, Yurovsky, Marchman, & Frank, 2019; Fourtassi, Bian, & Frank, 2018), it is not known how well the distinction between adjacent and non-adjacent contextual information would generalize cross-linguistically.

Age of acquisition norms provide a unique opportunity to investigate how children's environment supports their acquisition of knowledge and competence. The current study demonstrates strong and detailed links between words' usage patterns and AoA. A challenge for future research is to integrate these results with studies of individual differences, computational learning theory, and laboratory word learning studies. Under this vision, the goal is to integrate microlevel causal, cognitive, and neural explanations for word learning with observed developmental trajectories within natural environments. If successful, we believe this would serve as a model for how to explain ontogenesis in complex systems generally.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgments

This research was supported by the National Science Foundation (NSF) Graduate Research Fellowship under grant no. DGE-1144086, NSF grant SES-0527756 to G. Deák and J. Triesch, and by the Temporal Dynamics of Learning Center (NSF SMA-1005256).

Open Research badges

This article has earned Open Data badge. Data is available at www.github.com/CogDevLabUCSD/distributional-features-aoa.

Note

1. All analyses can be reproduced using the code and data at: www.github.com/CogDevLabUCSD/distributional-features-aoa.

References

Adamson, L. B., & Bakeman, R. (1991). The development of shared attention during infancy. In R. Vasta (Ed.), *Annals of child development* (Vol. 8, pp. 1–41). London: Jessica Kingsley Press.

Arnon, I., & Clark, E. V. (2011). Why brush your teeth is better than teeth–Children's word production is facilitated in familiar sentence-frames. *Language Learning and Development*, *7*(2), 107–129.

Barclay, J. R., Bransford, J. D., Franks, J. J., McCarreli, N. S., & Nitsch, K. (1974). Comprehension and semantic flexibility. *Journal of Verbal Learning and Verbal Behavior*, *13*, 471–481.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., Reilly, J., & Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*(1), 85–123.

Beckage, N., & Colunga, E. (2013). Using the words toddlers know now to predict the words they will learn next. In M. Knauff, M. Pauen, N. Sebanz & I. Wachsmuth (Eds.). *Proceedings of the 35th annual meeting of the Cognitive Science Society*. (pp. 163–168). Austin, TX: Cognitive Science Society.

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, *114*(49), 12916–12921.

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition*, *127*(3), 391–397.

Bhatt, R. S., Wilk, A., Hill, D., & Rovee-Collier, C. (2004). Correlated attributes and categorization in the first half-year of life. *Developmental Psychobiology*, *44*(2), 103–115.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2016). From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. In A. Papafragou, D. Grodner, D. Mirman, & J.C. Trueswell (Eds.). *Proceedings of the 38th annual conference of the Cognitive Science Society* (pp. 1691–1696). Austin, TX: Cognitive Science Society

Braginsky, M., Yurovsky, D., Marchman, V., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, *3*, 52–67

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–B44.

Brysbaert, M., Warriner, A., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27*(6), 843–873.

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, *110*(28), 11278–11283.

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J. (1995). A cross-linguistic study of early lexical development. *Cognitive Development*, *10*(2), 159–199.

Chang, L., & Deák, G. (2019). Maternal discourse continuity and infants' actions organize 12-month-olds' language exposure during object play. *Developmental Science*, *22*(3), e12770.

Childers, J. B., & Tomasello, M. (2002). Two-year-olds learn novel nouns, verbs, and conventional actions from massed or distributed exposures. *Developmental Psychology*, *38*(6), 967.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Clair, M. C. S., Monaghan, P., & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*, *116*(3), 341–360.

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372* (1711), 20160055.

de Barbaro, K., Johnson, C. M., Forster, D., & Deák, G. O. (2016). Sensorimotor decoupling contributes to triadic attention: A longitudinal investigation of mother–infant–object interactions. *Child Development*, *87* (2), 494–512.

Deák, G. O. (2000). Hunting the fox of word learning: Why "constraints" fail to capture it. *Developmental Review*, *20*(1), 29–80.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, *38* (1), 188–230.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, *37*, 66–108.

Eriksson, M., Westerlund, M., & Berglund, E. (2002). A screening version of the Swedish communicative development inventories designed for use with 18-month-old children. *Journal of Speech, Language, and Hearing Research*, *45*, 948–960.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the MacArthur communicative development inventories at ages one and two years. *Child Development*, *71*, 310–322.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., & Reznick, J. S. (2007). *MacArthur-Bates communicative development inventories: User's guide and technical manual*. Baltimore, MD: Brookes.

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology*, *27*(2), 209–221.

Fourtassi, A., Bian, Y., & Frank, M. C. (2018). Word learning as network growth: A cross-linguistic analysis. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.). *Proceedings of the 40th annual meeting of the Cognitive Science Society* (pp. 384–389). Austin, TX: Cognitive Science Society.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*(1), 1–24.

Frermann, L., & Frank, M. C. (2017). Prosodic features from large corpora of child-directed speech as predictors of the age of acquisition of words. arXiv preprint arXiv:1709.09443.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3–55.

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology*, *69*(2), 133–149.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, *7*(5), 219–224.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Horowitz, A. C., & Frank, M. C. (2015). Young children's developing sensitivity to discourse continuity as a cue for inferring reference. *Journal of Experimental Child Psychology*, *129*, 84–97.

Huebner, P. A., & Willits, J. A. (2018). Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*, *9*, 133.

Kauschke, C., & Hofmeister, C. (2002). Early lexical development in German: A study on vocabulary growth and vocabulary composition during the second and third year of life. *Journal of Child Language*, *29*(4), 735–757.

Laakso, A., & Smith, L. B. (2007). Pronouns and verbs in adult speech to children: A corpus analysis. *Journal of Child Language*, *34*(4), 725–763.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321.

Lieven, E. V., Pine, J. M., & Barnes, H. D. (1992). Individual differences in early vocabulary development: Redefining the referential-expressive distinction. *Journal of Child Language*, *19*(2), 287–310.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk, Volume II: The database*. New York, NY: Psychology Press.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*(1), 57–77.

Mason, G. M., Kirkpatrick, F., Schwade, J. A., & Goldstein, M. H. (2018). The role of dyadic coordination in organizing visual attention in 5-month-old infants. *Infancy*, *24*(2), 162–186.

McGillion, M. L., Herbert, J. S., Pine, J. M., Keren-Portnoy, T., Vihman, M. M., & Matthews, D. E. (2013). Supporting early vocabulary development: What sort of responsiveness matters? *IEEE Transactions on Autonomous Mental Development*, *5*(3), 240–248.

Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology*, *20*(2), 158–190.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014–9019.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (Eds.). *Advances in neural information processing systems* (pp. 3111–3119). Cambridge, MA: MIT Press.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, *96*(2), 143–182.

Morford, J. P., & Goldin-Meadow, S. (1997). From here and now to there and then: The development of displaced reference in homesign and English. *Child Development*, *68*(3), 420–435.

Murphy, C. M. (1978). Pointing in the context of a shared activity. *Child Development*, *49*(2), 371–380.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11–28.

Oliver, B., Dale, P. S., Saudino, K. J., Petrill, S. A., Pike, A., & Plomin, R. (2003). The validity of a parent-based assessment of cognitive abilities in three-year olds. *Early Child Development and Care*, *172*, 337–348.

Ororbia, I. I., Alexander, G., Mikolov, T., & Reitter, D. (2017). Learning simpler language models with the delta recurrent neural network framework. arXiv preprint arXiv:1703.08864.

Phillips, L., & Pearl, L. (2015). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science*, *39*(8), 1824–1854.

Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2004). An improved method for deriving word meaning from lexical co-occurrence. *Cognitive Psychology*, *7*, 573–605.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906–914.

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child development*, *83*(5), 1762–1774.

Rowe, M. L. (2013). Decontextualized language input and preschoolers' vocabulary development. *Seminars in Speech and Language*, *34*(4), 260–266.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663–12668.

Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent-child discourse. *Children's Language*, *4*, 1–28.

Sahni, S. D., Seidenberg, M. S., & Saffran, J. R. (2010). Connecting cues: Overlapping regularities support cue discovery in infancy. *Child Development*, *81*(3), 727–736.

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word–referent learning. *Trends in Cognitive Sciences*, *18*(5), 251–258.

Smith, L. B., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, *16*(3), 407–419.

Sullivan, J., & Barner, D. (2016). Discourse bootstrapping: Preschoolers use linguistic discourse to learn new words. *Developmental Science*, *19*(1), 63–75.

Swingley, D., & Humphrey, C. (2018). Quantitative linguistic predictors of infants' learning of specific English words. *Child Development*, *89*(4), 1247–1267.

Tamis-LeMonda, C. S., Kuchirko, Y., & Song, L. (2014). Why is infant language learning facilitated by parental responsiveness? *Current Directions in Psychological Science*, *23*(2), 121–126.

Tomasello, M. (2000). The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences*, *4*(4), 156–163.

Tomasello, M., & Mervis, C. B. (1994). The instrument is great, but measuring comprehension is still a problem. *Monographs of the Society for Research in Child Development*, *59*, 174–179.

Trueswell, J. C., Lin, Y., Armstrong III, B., Cartmill, E. A., Goldin-Meadow, S., & Gleitman, L. R. (2016). Perceiving referential intent: Dynamics of reference in natural parent-child interactions. *Cognition*, *148*, 117–135.

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*(3), 375–382.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13–15), 2149–2165.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–262.

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental science*, *16*(6), 959–966.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Table S1. RMSE for different values of minimum frame/context word frequency.

Table S2. RMSE for different numbers of PCs.

Table S3. RMSE for different winsorization thresholds.

Table S4. Variance Inflation Factors for distributional features.

Table S5. Variance Inflation Factors for baseline features.